ISRN TELCORDIA--2011-04+PR-0GARAU
Prepared for: Office of Naval Research

Fast Multiscale Algorithms for Information Representation and Fusion

Technical Progress Report No. 4

Devasis Bassu, Principal Investigator

Contract: N00014-10-C-0176

Telcordia Technologies

One Telcordia Drive

Piscataway, NJ 08854-4157

July 2011

Report Documentation Page	*Form Approved OMB No. 0704-0188*

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **JUL 2011**	2. REPORT TYPE	3. DATES COVERED **00-00-2011 to 00-00-2011**
4. TITLE AND SUBTITLE **Fast Multiscale Algorithms For Information Representation And Fusion**		5a. CONTRACT NUMBER
		5b. GRANT NUMBER
		5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)		5d. PROJECT NUMBER
		5e. TASK NUMBER
		5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Telcordia Technologies,One Telcordia Drive,Piscataway,NJ,08854**		8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)		10. SPONSOR/MONITOR'S ACRONYM(S)
		11. SPONSOR/MONITOR'S REPORT NUMBER(S)
12. DISTRIBUTION/AVAILABILITY STATEMENT **Approved for public release; distribution unlimited**		
13. SUPPLEMENTARY NOTES		
14. ABSTRACT		
15. SUBJECT TERMS		

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**12**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

1 Abstract

In the fourth quarter of the work effort, we focused on a) fine tuning and bug fixes for the randomized SVD and ANN algorithms, b) initial selection of real-world data sets/problems and applications using the developed algorithms, and c) preliminary design of the Multiscale Singular Value Decomposition (SVD) algorithm. This report presents motivation and a rough design sketch for the new multiscale SVD algorithm along with details on the selected data sets and possible applications.

The project is currently on track – in the upcoming quarter, we will continue applying the developed algorithms to various data sets and advance the design of the multiscale SVD algorithm. Also, we expect to provide an open-source home for the randomized SVD and ANN algorithms. No problems are currently anticipated.

Table of Contents

1 ABSTRACT II

2 SUMMARY 1

3 INTRODUCTION 2

4 METHODS, ASSUMPTIONS AND PROCEDURES 3

4.1 Multiscale Singular Value Decomposition 3

4.1.1 Singular Value Decomposition 3

4.1.2 Motivation 3

4.1.3 Rough Sketch of Algorithm 4

4.2 Deliverables / Milestones 5

5 RESULTS AND DISCUSSION 6

5.1 Application: IP Traffic Profile Analysis 6

5.2 Application: Line-of-Sight Determination 6

5.3 Application: Text Retrieval / Indexing 6

6 CONCLUSIONS 8

7 REFERENCES 9

2 Summary

In this quarter, we performed fine-tuning and bug fixes for the randomized SVD and ANN algorithms. We are also developing convenient command-line invocation tools in addition to the previously developed APIs. Various real-world data sets/applications were selected for trying out the developed algorithms. Algorithm design work was started for the new multiscale SVD algorithm.

The project is currently on track – in the upcoming quarter, we will continue applying the developed algorithms to various data sets and design of the multiscale SVD algorithm. No problems are currently anticipated.

3 Introduction

The primary project effort over the last quarter focused on bug fixes and fine-tuning the randomized SVD and ANN algorithms [1][2]. In addition to extending the auto-regression software test suite, we are developing convenient command-line tools to invoke the developed algorithms. Various real-world data sets/applications were selected for trying out the developed algorithms (see Section 5). Finally, we started work on the design of the new multiscale Singular Value Decomposition algorithm [3][4][5]. Motivation along with a rough sketch of the algorithm is provided in Section 4.

We have started the process for finding an open-source home for the software developed during the course of this program. We expect to start transitioning the developed algorithms to their new open-source home in the upcoming quarter.

4 Methods, Assumptions and Procedures

4.1 Multiscale Singular Value Decomposition

The following describes the new multiscale Singular Value Decomposition algorithm and provides a preliminary sketch of the algorithm design.

We start with the definition of the standard Singular Value Decomposition (SVD) algorithm from linear algebra.

4.1.1 Singular Value Decomposition

Given an $m \times n$ matrix A of rank $k < \min(m,n)$, the SVD represents A in the form

$$A = U \circ D \circ V^*$$

where D is a $k \times k$ diagonal matrix whose elements are non-negative, and U and V are matrices (of sizes $m \times k$ and $n \times k$, respectively) whose columns are orthonormal. The compression provided by the SVD is optimal in terms of accuracy, and has a simple geometric interpretation: it expresses each of the columns of A as a linear combination of the k (orthonormal) columns of U; it also represents the rows of A as linear combinations of (orthonormal) rows of V; and the matrices U, V are chosen in such a manner that the rows of U are images (up to a scaling) under A of the columns of V.

4.1.2 Motivation

The SVD provides a fundamental decomposition of any given matrix (from a data analysis perspective, one way to think of a matrix would be a stacked finite set of d-dimensional data points). The decomposition is optimal assuming that the underlying geometry of the points is linear which may however not be the case. Also, the decomposition is global in the sense that it takes all the points into account – what this means is that for large data sets it provides a linear approximation to the geometry at the global scale; the computed linear basis may or may not be optimal or even appropriate for subsets of the larger data set at smaller scale sizes (zoomed in). Figure 1 provides an example of this phenomenon. The top-left figure shows the original data set comprising of three clusters of points with different geometries along with the first two principal axes (first two singular vectors of the SVD). The remaining three insets in Figure 1 show the first two principal axes computed for each of the clusters in the original data set. Observe that the local geometries are quite different from the global geometry. Using the SVD basis to represent the data set is clearly going to be sub-optimal for analysis at a smaller scale size.

The multiscale SVD provides a multiscale representation of the data set which captures local geometries. This should also provide good representations for the case of global data sets with non-linear structures possessing locally linear geometries. A rough sketch of the algorithm is provided next.

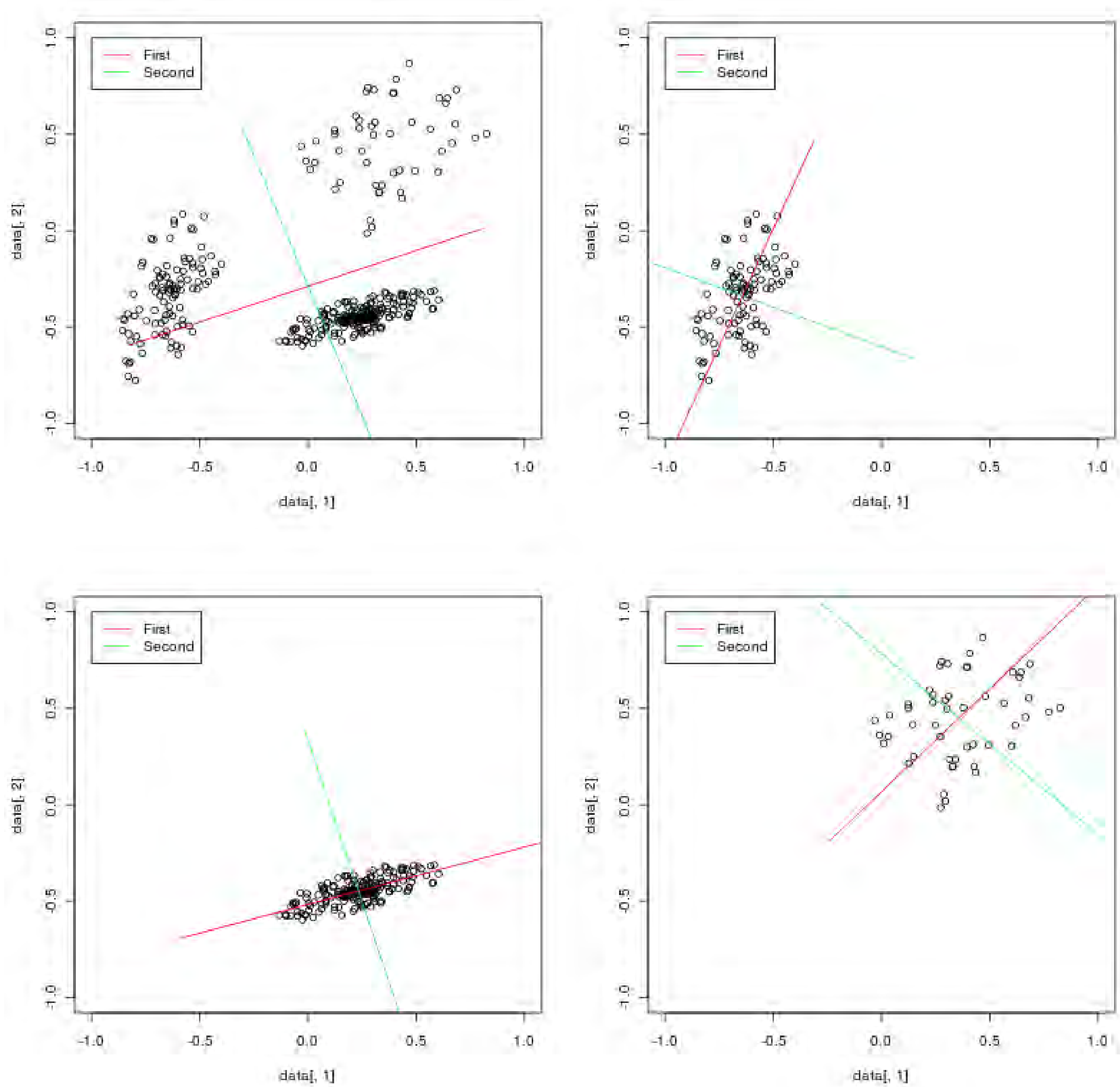

Figure 1: First two principal axes for a) original data set comprising different geometries (top-left); b) three individual sub-sets of the original data set

4.1.3 Rough Sketch of Algorithm

The central idea is to recursively partition the given point set into smaller bins and compute the SVD for each bin. One possibility is to compute the first singular vector the point set and then use that to split the point set into two sets. Next, compute the first singular vector for each of the smaller point sets. Continue recursively. The collection of singular vectors for any branch of the point set tree provides a representation of the points in that set.

4.2 Deliverables / Milestones

Date	Deliverables / Milestones	Status
Oct 2010	Progress report for period 1, 1st quarter	✓
Jan 2011	Progress report for period 1, 2nd quarter / complete randomized matrix decompositions task	✓
Apr 2011	Progress report for period 1, 3rd quarter / complete approximate nearest neighbors task	✓
Jul 2011	Progress report for period 1, 4th quarter / complete experiments – part 1	✓
Oct 2011	Progress report for period 2, 1st quarter	
Jan 2012	Progress report for period 2, 2nd quarter / complete multiscale SVD task	
Apr 2012	Progress report for period 2, 3rd quarter	
Jul 2012	Progress report for period 2, 4th quarter / complete experiments – part 2	
Oct 2012	Progress report for period 3, 1st quarter	
Jan 2013	Progress report for period 3, 2nd quarter / complete multiscale Heat Kernel task	
Apr 2013	Progress report for period 3, 3rd quarter	
Jul 2013	Final project report + software + documentation on CDROM / complete experiments – part 3	

The next section provides details about the selected real-world data sets and applications.

5 Results and Discussion

We present details of the selected real-world data sets and potential applications. The primary considerations for the selection process were that a) the data sets are large and reflect real-world dynamics, b) the developed algorithms could be used to perform analysis on them, and c) some amount of ground truth is available to verify/validate the results.

5.1 Application: IP Traffic Profile Analysis

This data set comprises IP traffic collected at a single CISCO switch within the Applied Research network at Telcordia. The data was collected using nfdump [6] and stored in NetFlow [7] record format. The data spans the time period starting from 03-Aug-2009 15:00 to 13-Mar-2010 17:00 resulting in over 200GB of data. A sample NetFlow record is provided below.

```
Date flow start          Duration Proto   Src IP Addr:Port       Dst IP Addr:Port       Flags Tos  Packets Bytes Flows
2005-08-30 06:53:53.370  63.545 TCP  113.138.32.152:25   ->  222.33.70.124:3575  .AP.SF  0     62    3512    1
```

The objective is to build profiles for each local IP on the network associated with the switch along with high-level operational profiles for categorizing the IP's (e.g., weekly profiles, holiday/workday profiles, desktop/server profiles). Subsequently, the profiles will be used to predict/classify new/unknown data for any given IP. A known complication is that the IPs are not all statically allocated and may have been reused by different machines during the course of the data collection (list of static addresses can be obtained easily).

5.2 Application: Line-of-Sight Determination

The objective here is to answer if one can determine if a received signal consists of signal+multipath, or direct signal only using sampled RF signal measurements. Measurements are performed by placing a signal source in a known location and driving a route with several types of multipath conditions. This knowledge is important in geolocation applications where knowing whether a received signal is line-of-sight or not is necessary for the algorithms to work

This data set was collected on-site at the Telcordia Navesink campus at Red Bank, NJ. Two sets of DRS 9144(receiver) /9475(digitizer) pairs were used along with GPS receivers/loggers in addition to signal generators and antennas to generate and collect the RF data. Each different multipath condition (full/partial/zero line-of-sight) along the route was time-stamped and recorded. The data set is around 25GB of GPS time-stamped raw baseband I + Q measurements.

5.3 Application: Text Retrieval / Indexing

We have access to a large number of public textual corpora at Telcordia (from previous work efforts on Latent Semantic Indexing) covering a large number of domains including scientific documents, UN meeting transcripts, movie reviews, legal and religious documents. Additionally, we have downloaded Twitter data [8] available from http://ckan.net/package/twitter-social-graph-www2010. This data set comprises results of a full crawl of the entire Twitter site with 41.7 million user profiles, 1.47 billion social relations, 4,262 trending topics, and 106 million tweets.

The objective with the social network data would be to form profiles of users, identify topics and groups. Fast indexing and retrieval would be associated tasks with all text data sets.

6 Conclusions

The project is on track with wrapping up development of the randomized SVD and ANN algorithms along with an early start on the design of the multiscale SVD algorithm. We will continue experimenting on the selected real-world data sets using the developed algorithms in the next quarter.

No problems are currently anticipated.

7 References

[1] V. Rokhlin, M. Tygert, *A fast Randomized Algorithm for the Overdetermined Linear Least Squares Regression*, PNAS, vol. 105, No. 36, pages 13212-13217, 2008.

[2] P. Jones, A. Osipov, V. Rokhlin, *A Randomized Approximate Nearest Neighbors Algorithm*, Research Report YALEU/DCS/RR-1434, Yale University, September 14, 2010

[3] G. Lerman, *Quanitfying curvelike structures of measures by using Jones quantities*, C.P.A.M., vol. 56, issue 8, pages 1294-1365.

[4] P. Jones, *Square functions, Cauchy integrals, analytic capacity, and harmonic measure*, Harmonic Analysis and Partial Differential Equations, Springer Lecture Notes in Math. No. 1384 (1989), pages 24-68.

[5] P. Jones, *Rectifiable sets and the Traveling Salesman Problem*, Inventiones Math. 102 (1990), pages 1-15.

[6] nfdump, URL: http://nfdump.sourceforge.net/

[7] NetFlow, URL:http://en.wikipedia.org/wiki/Netflow

[8] H. Kwak, C. Lee, H. Park, S. Moon, *What is Twitter, a social network or a news media?*, Proceeding of the 19th international conference on WWW'10, ACM, April 2010.